Development of SRI's 1997 Broadcast News Transcription System*

Ananth Sankar, Fuliang Weng, Ze'ev Rivlin, Andreas Stolcke, and Ramana Rao Gadde

Speech Technology and Research Laboratory
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025

ABSTRACT

This paper describes SRI's 1997 broadcast news transcription system used for the 1997 DARPA H4 evaluations. Our system had several novel components. These include automatic segmentation of entire broadcast shows, word-internal and crossword acoustic models robustly estimated with a new Gaussian Merging-Splitting (GMS) algorithm, the use of trigram language models (LMs) in lattices instead of for rescoring N-best lists, and an LM pruning algorithm that allows efficient representation of high-order (like 4- or 5-gram) LMs. We briefly describe these features and give comparative experimental results. We achieved a 18.7% relative improvement in performance on our 1996 H4 partitioned evaluation (PE) development test set as compared to our 1996 H4 PE evaluation system.

1. Introduction

In recent years there has been increasing interest in developing large-vocabulary continuous speech recognition (LVCSR) algorithms for speech found in real sources such as telephone conversations and broadcast news. Broadcast news, in particular, has been the test bed for the DARPA continuous speech recognition (CSR) evaluations over the last few years, and represents a significant challenge to speech recognition researchers.

Many interesting problems are associated with the automatic recognition of broadcast news. One problem is that the speech is in the form of a single long stream, whereas typical automatic speech recognition (ASR) systems are designed to process sentence-length units of speech. ASR systems work best when the segment to be recognized is homogeneous with respect to speaker and acoustic condition. It is also desirable, both for ASR and for speech understanding, for the segments to correspond to linguistic units such as sentences or phrases. An interesting challenge is to develop algorithms that can automatically segment a long stream of speech according to such criteria. Another problem with broadcast news is the many different variabilities of speech, such as conversational speech, noisy speech, speech in the presence of music, non-native speech, or a combination of these variations. In addition, these variations are constantly changing from one to another, making the ASR problem very challenging. Since automatic recognition of broadcast news can be used for applications such as information archiving and retrieving, focusing effort on this task serves the dual purpose of improving technology and opening up interesting LVCSR applications.

In this paper we describe SRI's 1997 broadcast news transcription system. We start with an overview of the system, and then describe the individual components. Comparative experimental results are given to show the benefit of new approaches. The techniques we developed in 1997 resulted in a 18.7% improvement in performance on our 1996 H4 partitioned evaluation (PE) development test set as compared to our 1996 H4 PE evaluation system.

2. Overview of SRI's H4 System

The 1997 DARPA H4 evaluation test data was in the form of a single 3-hour-long file. Different broadcast shows, or segments of these shows, were spliced together to create this file. We processed the data in the file by using our broadcast news transcription system as follows:

1. **Acoustic Segmentation**: The 3-hour waveform was automatically segmented by recognizing the waveform with a fast context-dependent (CD) phonetically tied mixture (PTM) parallel male/female recognition system, and then segmenting at regions where the system hypothesized background or silence. In addition, we segmented at all gender changes located by the male/female recognition system. The segments were clustered into acoustically similar groups using bottom-up agglomerative clustering.

2. **First Pass Recognition**: The segments were recognized using gender-dependent, speaker-independent (SI) Genonic hidden Markov models HMMs [1] trained using a new Gaussian Merging-Splitting (GMS) algorithm [2], and the 100 hours of H4 training data. This recognition was done with a 48,000-word bigram language model.

3. **Adapting Models**: The data from each segment cluster was used along with the hypotheses generated in Step 2 to adapt the gender-dependent HMMs to each test

*This work was sponsored by DARPA through the Naval Command and Control Ocean Surveillance Center under contracts N66001-94-C-6048 and N66001-94-C-6046.

segment cluster. We used maximum-likelihood (ML) transformation-based adaptation [3, 4], with a block-diagonal affine matrix transformation of the HMM mean vectors [5].

4. **Trigram Lattice Generation**: The test-segment-cluster adapted models were used to create bigram lattices, which were then expanded to trigram lattices.

5. **Recognition from Trigram Lattices:** The gender-dependent SI models were used to run recognition for all the segments using the trigram lattices. Ideally, we would have used the test-segment-cluster adapted models during this stage, but for logistic reasons we did not do this during the evaluations. Since trigram lattices are used, the hypotheses in this step are significantly better than those generated in Step 2.

6. **Adapting Models and Creating N-Best Lists**: The recognition hypotheses from Step 5 were used to adapt models for each of the test segment clusters. Both a block diagonal affine matrix transform of the HMM means and a variance scaling transform [4, 5] were used in this step. The adapted models were then used to run forward-backward recognition on the trigram lattices to create word-dependent n-best lists.

7. **Adapting Crossword Models**: The hypotheses from Step 5 were used to adapt crossword models which were trained using the GMS algorithm. Again, both mean and variance adaptation were used, as in Step 6.

8. **Rescoring N-Best Lists**: The n-best lists generated in Step 6 were rescored using the test-segment-cluster adapted crossword models generated in Step 7. The n-best lists were also rescored with a 48,000-word vocabulary 5-gram LM.

9. **Combining Scores**: Finally, the scores from four different knowledge sources were linearly combined to give the final hypotheses. The knowledge sources were

   (a) Non-crossword test-segment-cluster adapted Genonic HMMs

   (b) Crossword test-segment-cluster adapted Genonic HMMs

   (c) 48,000-word vocabulary 5-gram LM

   (d) Number of words in hypotheses (used to penalize word insertions)

To measure performance during development of our system, we used the 1996 H4 PE development test set. In the PE test, the speech is hand-segmented into segments homogeneous with respect to speaker and acoustic condition. Since the 1997 H4 evaluation was an unpartitioned evaluation (UE), where no hand-generated segments were given, there is a mismatch between the data we used for development and the final evaluation data. However, because only the segmentation step is particular to the UE, we believe that we can get a good estimate of the performance of our system by using the 1996 PE development test set. In particular, the word error rate for the 1997 system was 26.1% as compared to 32.1% with our 1996 system, which is a 18.7% relative improvement.

3. Feature Extraction

The front-end feature extraction was based on mel-frequency cepstrum processing. The speech was hamming-windowed, with a 25.6ms window advanced every 10 ms. Each frame was represented by 12 mel-frequency cepstrum coefficients, the log energy, and their first- and second-order time derivatives (delta and delta-delta features), for a resulting 39-dimensional feature vector.

4. Vocabulary and Dictionary

A 48,000-word vocabulary was selected by choosing the most frequent words from the 1996 H4 language model (LM) training texts and adding all words that occurred at least twice in the 1996 acoustic training transcripts. This vocabulary resulted in an out-of-vocabulary (OOV) rate of 0.9% on the 1996 H4 development test set. We used version 0.3 of the CMU dictionary modified at SRI to make sure that pronunciations existed for all the 48,000 vocabulary words.

5. Acoustic Segmentation and Clustering

Our acoustic segmentation algorithm is a modification of the algorithm we used in 1997 to segment long PE segments [6]. In the 1996 PE data, each stream was guaranteed to contain only speech, and to come from a single speaker. Thus, the problem was simply to chop the stream into shorter segments, which was done by segmenting at non-speech regions hypothesized by Viterbi beam search (with a low pruning beamwidth for fast recognition) using a gender-independent CD PTM recognition system.

For the UE data, the stream contains speech and also many long non-speech regions. In addition, there are no given speaker boundaries. We modified our previous algorithm to use a parallel male/female CD PTM system for recognition, and included 5 minutes of non-speech data from the H4 acoustic training data to train the non-speech model. The segmentation algorithm was modified to remove any non-speech segments longer than 1 second, and then chop at the remaining non-speech segments to create nominally 10-second segments. In addition, a new segment was created whenever a gender change occurred. The resulting segments are thus nominally 10 seconds long, and are labeled by gender.

We tested the segmentation algorithm with the 1996 H4 de-

Segment Type	Models	
	SI	Cluster-Adapted
PE	37.9	35.5
UE	39.4	37.6

Table 1: Word error rates (%) for the PE and UE segments

Models	Word Error (%)
Condition-Specific	41.12
Single H4 Model	38.61

Table 2: Comparison of condition-specific models vs. a single H4 model

velopment test data. For four broadcast shows, both PE and UE index files were provided. We ran recognition on the PE and UE segments for these shows using a 20,000-word bigram LM, and non-crossword gender-dependent Genonic HMMs. Both SI models and segment-cluster-adapted models were used. Table 1 gives word error rates for the PE and UE recognition runs. For the SI models, the word error rate was 1.5% (absolute) worse for the UE than for the PE. However, for the adapted models this difference increased to 2%, possibly because a single UE segment may contain speech from multiple acoustic conditions or speakers, giving segment clusters that are not acoustically homogeneous, and thus degrading the adapted models.

The segments were clustered using bottom-up agglomerative clustering as in our 1996 system [6, 7]. However, we modified the way in which the distances were computed between the segments. In our previous work [6, 7], a separate Gaussian mixture model (GMM) was trained for each segment, and the distance between the segments was given by the symmetric relative entropy computed using these GMMs [6, 7]. Since some segments have very little data, it is difficult to estimate a full GMM for each segment. We modified our approach by training a single GMM for all the segments, and using a separate mixture weight distribution for each segment to these shared Gaussians. The distance between two segments is then defined as the weighted-by-counts increase in entropy of the mixture weight distribution due to clustering two segments. This is identical to the approach we use for HMM state clustering [1]. The performance of the new clustering algorithm was found to be slightly better than that of the approach we used last year.

6. Acoustic Modeling

For the 1997 broadcast news transcription system, we trained gender-dependent Genonic HMM models [1] using only the nominally 100 hours of H4 acoustic training data. This is a deviation from our 1996 system, where we adapted models, trained with the Wall Street Journal or Switchboard training data, to each of the seven individual acoustic focus conditions defined by the H4 evaluation committee [8], using the first 50 hours of acoustic training data. This creates seven condition-specific models for each gender. In 1996, we participated in the PE test, where the acoustic focus condition was given for each test segment. We believed that adapting models to each focus condition would give better performance than a single H4 model since the models would be tuned to the specific focus conditions. After the 1996 evaluations, we trained a single gender-dependent H4 model using the first 50 hours of training data. This approach was taken by BBN in the 1996 evaluations [9]. Table 2 gives recognition word error rates with the 20,000 word bigram LM we used for the 1996 evaluations for the male subset of the 1996 development test set. The single H4 model gave a relative 6.1% lower word error rate than the condition-specific models. Since training a single H4 model is also easier, we chose to use this approach for our 1997 broadcast news system.

We have recently developed the GMS algorithm to train state-clustered HMMs. We use Genonic HMMs [1], where each HMM state cluster shares the same set of Gaussians (or Genone), and a separate mixture weight is used for each state. The GMS algorithm uses iterative Gaussian splitting and training to generate the required number of Gaussians per Genone. At each stage of training the Gaussians are iteratively merged until all Gaussians have at least a threshold of data. For the HMM parameters, this technique was found to give more robust estimates than our previous training algorithm. The GMS algorithm is described elsewhere in these proceedings [2].

We used the GMS algorithm to train a separate non-crossword H4 model for each gender. We used 7761 triphones for the males and 6723 triphones for the females. Since the GMS algorithm guarantees robust parameter estimation, we explored HMM structures with a very large number of Gaussians per Genone (and fewer Genones) as compared to what we used previously. In particular, we used 535 Genones for the males and 569 Genones for the females. For both cases, we used 128 Gaussians per Genone. Based on experiments with the GMS algorithm, we have some evidence to support the hypothesis that for a fixed number of Gaussians, better performance is achieved by using fewer Genones and more Gaussians per Genone as compared to our previous models where we used more Genones and fewer Gaussians per Genone. An explanation for this is given in [2].

We used Genonic crossword models to rescore n-best lists. The triphones in the crossword models are word-position in-

SI	Adapted
31.78	29.97

Table 3: Comparison of SI and mean-adapted models

dependent. There were 16,728 triphones for the males and 13,368 triphones for the females. Due to a lack of time, we did not experimentally select the HMM structure for the crossword HMMs, but decided to use about 2000 Genones and 32 Gaussians per Genone.

7. Test-Segment-Cluster Adaptation

As in our 1996 system, we adapted the SI HMMs to each test-segment-cluster by using unsupervised transcription-mode ML transformation-based adaptation [3, 4]. The transformations were a block diagonal affine matrix transform of the HMM mean vectors [5], and a scaling transform for the variance vector [4, 5]. We used three separate transforms, one of them being tied to the non-speech (silence) model. The other transforms are tied to phone classes determined by a human expert. Table 3 shows the performance gain from mean adaptation for the 1996 H4 development test set. For these runs, we used trigram lattices, which we recently implemented [10]. From the table, we see a relative 5.7% improvement from using adaptation. This is less than the 8.3% improvement we reported using last year's system [6]. This could be explained by the fact that our SI models have improved over last year. Thus, the further improvement possible from adaptation may decrease.

We also experimented with iterative adaptation. We tried two different approaches. In the first, we used the hypotheses generated by the previously adapted models to re-adapt the models iteratively for five iterations. At each stage we use a constant number of three transforms. In the second approach, we started with a single transform, increasing it to two, three, six, and eleven transforms in subsequent iterations. In addition, in this approach, adaptation is stopped if the transcriptions do not change from one iteration to the next [11]. Table 4 shows that we did not achieve any improvement by using iterative adaptation, and hence we did not use it in our final system.

Number of Transforms	SI	Adapted Models				
		1	2	3	4	5
Fixed at 3	31.8	30.0	29.9	29.9	29.9	29.8
Variable	31.8	30.0	30.0	30.1	30.1	30.1

Table 4: Effect of iterative adaptation

8. Trigram Lattice Generation

In previous years, we have used trigram and higher-order LMs to rescore n-best lists [12, 13, 6]. However, it is well known that trigram LMs give a drastic improvement over bigram LMs. Thus, it makes sense to use them earlier in the search. This year, we developed new lattice-based search capability by implementing a new bigram lattice algorithm and algorithms to expand these to trigram lattices [10]. We achieved about 5% relative improvement in performance on a male subset of the 1996 H4 development test set by running recognition from our new trigram lattices as compared to trigram rescoring of n-best lists generated using our 1996 H4 PE evaluation system.

9. LM Description

Three different LMs were used in the system. The first is a bigram LM trained using the 1996 H4 LM training corpus and the first 50 hours of H4 acoustic transcripts. This LM is used to create a word graph to run recognition in Step 2 of the system to get hypotheses for ML adaptation in Step 3. The trigram LM used in Step 4 for the trigram lattice expansion was trained using the 1996 H4 LM corpus, the first 50 hours of H4 acoustic transcripts, the 1995 H3 LM training corpus (which was drawn from North American Business News (NABN) texts), and the Switchboard-I training corpus. For the final rescoring LM used in Step 8, a 5-gram LM was estimated using the 1996 H4 LM corpus and the 1995 H4 LM training corpus (which was also drawn from NABN texts, but included non-financial data and a later cutoff date than the 1995 H3 training data), and a trigram LM was estimated for the first 50 hours of H4 acoustic transcripts and Switchboard-I. The 5-gram LM was pruned using a newly developed entropy-based pruning technique that drastically reduces the number of n-grams in the model without altering its performance [14].

In all the LMs, multiple corpora were used by training separate LMs for each and then interpolating the language models. The interpolation weights were estimated so as to minimize the perplexity on the 1996 development test transcriptions. All LMs used Katz backoff [15] and Good-Turing discounting.

10. N-best List Rescoring

The trigram lattices generated in Step 4 were used to generate n-best lists using the test-segment-cluster-adapted non-crossword models and the word-dependent n-best algorithm [16]. These n-best lists were then rescored with 5-gram LMs, test-segment-cluster-adapted crossword models, and a word insertion penalty. For crossword rescoring, the n-best lists were represented in the form of a tree lattice, resulting in very fast and memory-efficient rescoring. In addition to these three knowledge sources, we used the scores from the test-segment-cluster-adapted non-crossword models. The scores from each knowledge source were linearly combined, with the combination weights being found by a grid search to minimize

Knowledge Sources	1997 System	1996 System
	Choose highest combined score	
1996 trigram, non-crossword	29.0	32.5
1996 4-gram, non-crossword	28.4	32.1
1997 5-gram, non-crossword	28.3	-
1997 5-gram non-crossword, crossword	27.0	-
	Choose lowest expected word error	
1997 5-gram non-crossword, crossword	26.8	-

Table 5: Word error rates with different knowledge sources

word error on the 1996 PE development test set. Finally, the combined scores of the N-best hypotheses were normalized to estimate posterior probabilities for each hypothesis, which in turn were used to estimate expected word error counts for each hypothesis. The hypothesis with the lowest expected word error count was chosen to be the output of the recognition system [17].

Table 5 gives the word error rates using different knowledge sources. In all cases, we used word insertion penalty. The table gives the word errors for both the 1996 and 1997 systems. The difference in the 1996 and 1997 results for the first two rows is accounted for by the difference in the acoustic models used, and the new lattice-search strategies used in 1997. For all the results in the table, except the last row, we chose the best hypothesis by picking the one with the highest combined score. In the last row, we picked the hypothesis with the lowest expected word error count. The system we used in 1996 gave a 32.1% word error rate (with 4-gram LMs). Our 1997 system gave a word error rate of 26.8% resulting in a 16.5% relative improvement in performance. The system we used for the 1997 H4 evaluation was identical to that in the last row of Table 5 except that the 1997 evaluation system used automatic acoustic segmentation (Step 1 in the algorithm description given in Section 2). Our word error rate on the 1997 H4 evaluation data was 20.4%.

11. Retraining with Bug-Fixed Transcripts

After the evaluations, we noticed that there was a bug in the transcripts we used for training. Our transcripts contained no pause fillers in spite of their being present in the acoustics. On closer examination, we found that the NIST BN-filter, Version 1.11 was hard-wired to delete pause fillers from the transcripts. We had used the same scripts as we did in 1996,

Acoustic Condition	System	
	Evaluation	Bug fixed
F0	12.3	12.3
F1	28.6	27.3
F2	32.0	30.0
F3	31.8	32.0
F4	22.7	21.9
F5	20.2	20.5
FX	43.4	43.1
Total	26.8	26.1

Table 6: Word error rate (%) before and after bug fix for 1996 PE development test set

replacing the older BN-filter with the new version. We did not go through the process of verifying the resulting transcripts as our scripts had worked fine in 1996. Another bug we found was that we incorrectly mapped words from a small fraction of the sentences to a garbage model we used during training to segment OOV words.

We corrected these problems and retrained and reran our system on the 1996 H4 PE development set and the 1997 H4 UE evaluation set. Table 6 and 7 compare the performance of the evaluation system and the bug-fixed system across the different acoustic focus conditions for the 1996 H4 PE development test data and the 1997 H4 UE evaluation test data respectively. The word error rate on the 1996 development test set using the bug-fixed system was 26.1%. The word error rate on this test set using our 1996 H4 PE evaluation system was 32.1%. Thus we achieved a relative improvement of 18.7%. Part of this improvement can be attributed to the fact that we used nominally 50 hours of H4 acoustic training data to train the 1996 system, whereas we used nominally 100 hours to train the 1997 system. However, we observed a very modest improvement (about 1.5% relative) due to using the extra data, and most of the improvement can be attributed to the new techniques reported in this paper. Using the bug-fixed system, the error rate on the 1997 H4 evaluation test set was reduced from 20.4% to 20.0%.

12. Summary and Conclusion

For the 1997 SRI broadcast news transcription system, we developed and utilized several new techniques, including the GMS algorithm for HMM training, adapted crossword acoustic models, a new bigram lattice algorithm and trigram lattice expansion algorithm, and an algorithm to drastically prune LMs while maintaining their performance. As a result, we achieved an 18.7% relative improvement over our 1996 system.

Acoustic Condition	System	
	Evaluation	Bug fixed
F0	13.6	13.4
F1	20.5	20.1
F2	26.2	25.0
F3	32.4	33.2
F4	24.4	24.8
F5	28.1	28.0
FX	38.1	36.3
Total	20.4	20.0

Table 7: Word error rate (%) before and after bug fix for 1997 UE evaluation test set

References

1. V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 281–289, 1996.
2. A. Sankar, "Experiments with a Gaussian Merging-Splitting Algorithm for HMM training for Speech Recognition," in *Proceedings of DARPA Speech Recognition Workshop*, (Lansdowne, VA), February 1998.
3. V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.
4. A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 190–202, May 1996.
5. L. Neumeyer, A. Sankar, and V. Digalakis, "A Comparative Study of Speaker Adaptation Techniques," in *Proceedings of EUROSPEECH*, pp. 1127–1130, 1995.
6. A. Sankar, L. Heck, and A. Stolcke, "Acoustic Modeling for the SRI Hub4 Partitioned Evaluation Continuous Speech Recognition System," in *Proceedings of the 1997 DARPA Speech Recognition Workshop*, (Chantilly, VA), 1997.
7. L. Heck and A. Sankar, "Acoustic Clustering and Adaptation for Robust Speech Recognition," in *Proceedings of EUROSPEECH*, 1997.
8. R. Stern, "Specification of the 1996 Hub4 Broadcast News Evaluation," in *Proceedings of the DARPA Speech Recognition Workshop*, (Chantilly, VA), 1997.
9. F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, and J. Makhoul, "The 1996 BBN BYBLOS HUB-4 Transcription System," in *Proceedings of the 1997 DARPA Speech Recognition Workshop*, (Chantilly, VA), 1997.
10. F. Weng, A. Stolcke, and A. Sankar, "New developments in lattice-based search strategies in SRI's H4 system," in *Proceedings of DARPA Speech Recognition Workshop*, (Lansdowne, VA), February 1998.
11. M. J. F. Gales, "The Generation and Use of Regression Class Trees for MLLR Adaptation," Tech. Rep. CUED/F-INFENG/TR263, Cambridge University, 1996.
12. V. Digalakis, M. Weintraub, A. Sankar, H. Franco, L. Neumeyer, and H. Murveit, "Continuous Speech Dictation on ARPA's North American Business News Domain," in *Proceedings of the Spoken Language Systems Technology Workshop*, pp. 88–93, 1995.
13. A. Sankar, A. Stolcke, T. Chung, L. Neumeyer, M. Weintraub, H. Franco, and F. Beaufays, "Noise-resistant Feature Extraction and Model Training for Robust Speech Recognition," in *Proceedings of the 1996 DARPA CSR Workshop, Ardenhouse, NY*, 1996.
14. A. Stolcke, "Entropy-based Pruning of N-gram Backoff Language Models," in *Proceedings of DARPA Speech Recognition Workshop*, (Lansdowne, VA), February 1998.
15. S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 3, pp. 400–401, 1987.
16. R. Schwartz and Y.-L. Chow, "A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 701–704, 1991.
17. A. Stolcke, Y. Konig, and M. Weintraub, "Explicit Word Error Minimization In N-Best List Rescoring," in *Proceedings of EUROSPEECH*, pp. 163–166, 1997.